Genomic epidemiology of a major Mycobacterium tuberculosis outbreak: Retrospective cohort study in a low incidence setting using sparse time-series sampling

Folkvardsen, Dorte Bek; Norman, Anders; Andersen, Åse Bengård; Rasmussen, Erik Michael; Jelsbak, Lars; Lillebaek, Troels

Published in:
Journal of Infectious Diseases

Link to article, DOI:
10.1093/infdis/jix298

Publication date:
2017

Document Version
Peer reviewed version

Genomic epidemiology of a major *Mycobacterium tuberculosis* outbreak: Retrospective cohort study in a low incidence setting using sparse time-series sampling

Category: Major article

Running title: Genomic epidemiology of TB over decades

Dorte Bek Folkvardsen[#1], Anders Norman[#1,3], Åse Bengård Andersen[2], Erik Michael Rasmussen[1], Lars Jelsbak[*3], Troels Lillebaek[*1]

Summary: Important key information, such as mutation rate, core conserved SNPs and a possible historical origin, was derived from whole genome analysis of a big outbreak of tuberculosis using only a small fraction of isolates collected between 1992 and 2014.

[1]International Reference Laboratory of Mycobacteriology, Statens Serum Institut, DK-2300 Copenhagen, Denmark.

[2]Department of Infectious Diseases, Copenhagen University Hospital, Rigshospitalet, DK-2100 Copenhagen, and Research Unit for Infectious Diseases, Department of Clinical Research, University of Southern Denmark, Denmark.

[3]Department of Biotechnology and Biomedicine, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[#]Shared 1st author

*Shared last author, contributed equally.

Corresponding author:

Dorte Bek Folkvardsen, Phone: +4532683731, e-mail: dbe@ssi.dk

Word counts: Manuscript 3497, Abstract: 124

Abstract

Since 1992, Denmark has documented the largest outbreak of tuberculosis in Scandinavia ascribed to a single genotype, termed 'C2/1112-15'. As of spring 2017, the International Reference Laboratory of Mycobacteriology in Copenhagen has collected and identified isolates from more than a thousand cases belonging to this outbreak via routine MIRU-VNTR typing. Here, we present a retrospective analysis of the C2/1112-15 dataset, based on whole-genome data from a sparse time-series consisting of five randomly selected isolates from each of the 23 years. Even if these data are derived from only 12% of the collected isolates, we have been able to extract important key information, such as mutation rate, conserved single-nucleotide polymorphisms to identify discrete transmission chains, as well as the possible historical origins of the outbreak.

Introduction

Over the last two and a half decades, Denmark has experienced the largest sustained outbreak of tuberculosis (TB) in Scandinavia ascribed to a single genotype. Based on nationwide strain-typing of every *Mycobacterium tuberculosis* (*Mtb)* culture-positive case since 1992, it has been documented to progress at a surprisingly high rate [1–3]. Initially, the outbreak was identified by IS*6110* Restriction Fragment Length Polymorphism (IS6110-RFLP) genotyping and termed "Cluster 2" based on it being the second largest Danish epidemic cluster at the time[1]. Since, it has been identified through the now widely used 24 locus Mycobacterial Interspersed Repetitive Units – Variable Number of Tandem Repeats (MIRU-VNTR) typing method as the genotype 1112-15[4]. From 1992 through 2014, a total of 989 isolates belonging to C2/1112-15 (hereafter, just C2) were registered. Over the same period, the proportion of C2 increased from two (8/375) to fifteen percent of annually typed cases (56/365; own data). The majority of cases have been reported among Danish-born (DB) and Greenlandic-born (GB) males, especially within socially marginalized groups. Until recently, the outbreak was considered an isolated Danish problem, but transmission of the C2 genotype to Greenland[4] as well as neighboring countries (as of yet unpublished data) has since been observed.

A significant limitation with conventional strain genotyping is the lack of resolution to establish an exact chronology of transmission events. With the advent of whole-genome sequencing (WGS) and the increased availability of low-cost high-throughput sequencing platforms, the capacity to elucidate mutational landscapes and transmission has greatly enhanced the field of genomic epidemiology[5]. Although recent studies have used WGS to study the changes of *Mtb* during outbreaks[5–7], some even long-term continuous transmission[8], our knowledge about transmission and evolutionary dynamics of outbreaks over decades remain relatively limited. Here we report molecular characterization of the Danish C2 outbreak based on a representative time-series spanning the period 1992-2014.

Materials and Methods

Study population

The TB-incidence in Denmark pr. 100,000 over the selected period (1992-2014) ranged from 6-10 and was 7.1 by the end of 2014[9–11]. During the same period, a total of 9,501 Danish TB cases were culture-positive, from which 94% had at least one *Mtb* isolate successfully genotyped. Out of all typed isolates, 61% clustered with other cases [2,415 DB; 396 GB; 2,653 foreign-born (FB)], and 39% remained un-clustered [972 DB; 41 GB; 2481 FB]. Isolates from 989 cases [694 DB; 240 GB; 55 FB] were assigned to the C2 outbreak with either IS6110-RFLP or MIRU-VNTR typing, comprising 18% of all clustered Danish cases over the 23-year sampling period. The C2 cases were predominantly male (735/989) and from pulmonary TB (950/989).

Study design

In order to limit the number of isolates to be sequenced, we designed a sparse retrospective time-series, in which five isolates were picked from each year over the, which comprises 12% (115/989) of C2-typed isolates. Other than cases ideally being DB and from the Greater Copenhagen area, where C2 was first reported[1], isolates were picked at random.

Whole-genome sequencing and variant calling

DNA isolation and library preparation of *Mtb* isolates was performed as previously described[12]. Sequencing was conducted on Illumina MiSeq and NextSeq 500 platforms, using 2×150 bp paired-end chemistry. Libraries were diluted an extra four-fold when run on the Illumina NextSeq 500 sequencing platform. Paired-end reads were trimmed and quality filtered prior to variant calling as previously described[13]. Briefly, adapter fragments and low-quality reads were first removed with Trimmomatic[14]. Reads were then mapped to the *Mtb* H37Rv reference genome using BWA (http://bio-bwa.sourceforge.net/) and raw variants were called with samtools and bcftools (http://www.htslib.org/) using multi-sample pileups. Filtering criteria for VCF-formatted variants included a minimum read depth of 10 and a minimum variant coverage of at least 5 reads in at least one sample. Furthermore, variants were only considered at positions where the highest SNP-frequency exceeded 85%. Finally, all variants located in repetitive regions, such as *pe-*, *ppe-*, *pe_pgrs*-genes and transposons were removed. Sequences have been deposited in the European Nucleotide Archive under project accession number PRJEB20214.

Phylogenetic analysis

To compare with global *Mtb* SNPs, sequences from known representatives of *Mtb*-lineages 2-5, as well as members of lineage 4.8 were downloaded from the European Nucleotide Archive (**Table S1**) and variants were called in the same manner as described above. Multiple sequence alignments consisting of concatenated SNPs were generated from filtered vcf-files using an in-house perl script. Variants were assigned as missing at a given position if read coverage was less than five reads in that sample, and assigned as the reference if the SNP-frequency was below 70%. Positions with more than 5% missing bases were removed from the final alignment. Maximum likelihood trees were inferred with the program PHYML using the substitution model GTR, which consistently scored best with JMODELTEST.

Bayesian age estimation and phylogeny

We used BEAST (http://beast.bio.ed.ac.uk/) to infer Bayesian phylogeny and coalescent analysis on alignments consisting of concatenated SNPs. Briefly, clock rate and evolution model selection was performed as described previously[15]. The winning model (constant molecular clock with exponential population model) was run for 20 million iterations, with sampling every 1,000 states. To determine the evolutionary distance between the C2 and SAM5 sets (uncorrelated relaxed lognormal molecular clock with Bayesian

Skyline evolution model) the analysis ran for 50 million iterations. Monte Carlo Markov chain equilibrium and convergence of three independent chains was confirmed using TRACER (http://tree.bio.ed.ac.uk/software/tracer/). GENEIOUS (http://www.geneious.com/) was used to calculate consensus trees from trees generated in the first Markov chain, using a 70% support threshold and a 5% burn-in.

Results

MTBC lineage assignment and phylogenetic placement of C2 outbreak

We used the SNP-based nomenclature proposed by Coll et al[16] to assign the subset of C2 isolates to one of the currently defined MTBC sublineages. In all 115 isolates, we detected the GAC/GAT synonymous SNP at codon 51 of the Rv3417c gene, specific to MTBC sublineage 4.8 (L4.8), which extends from MTBC lineage 4 (Euro-American).

To determine phylogenetic relationships between C2 and related outbreaks in the literature, we screened publically available genomes for the L4.8 lineage-specific SNP-marker and retrieved a total of 254 genomes. Additionally, we retrieved 34 genomes from a global set of MTBC lineages 2-5, representing the three "modern" lineages 2-4[17], with a single lineage 5 isolate serving as an outgroup. Three of the included global reference isolates (L4_DY8, L4_N0109 and L4_V293AE) also contain the L4.8 marker.

From the resulting phylogeny, we observed that C2 constitutes a relatively deep-branching clade within L4.8 and is distinct from other outbreaks currently in the literature **(Figure 1a**). Upon closer inspection (**Figure 1b**), C2 extends from a diverse clade of isolates, all collected from Samara, Russia (2008-2010)[18].

To quantify clonal diversity, we then determined the number of pairwise SNPs between C2 and all other L4.8 isolates (**Figure 1c)**. The mean SNP distance from C2 to the rest of L4.8 was 293 SNPs, while the mean SNP distance from C2 to the H37Rv reference (L4.9) was 345. A small clade consisting of five Samara-isolates (hereafter: SAM5) stood out as being more closely related (mean distance: 154 SNPs) to C2 than all other L4.8 isolates.

We have recently described other isolates collected within the Danish Kingdom, including an outbreak in East Greenland (GE) collected 1992 – 2012 and a set of two linked Danish cases (Mu) from 1961 and 1994 that also harbor the L4.8 genotype[13,19]. However, neither study included 1112-15 MIRU-type isolates, and we observed quite a large distance between C2 and GE- and Mu (mean SNP distance: 313 and 294, respectively). Thus, consistent with the underlying genotyping data, we did not find genetic evidence linking these three studies.

C2 displayed a slightly bimodal distribution of pairwise SNP distances (**Figure 1d**), pointing to the presence of distinct clades. Based on the initial phylogenetic analysis, we defined the two clades C2-Mj (107 isolates) and C2-Mn (8 isolates). The mean within-clade minimum pairwise distance of C2-Mj and C2-Mn were 1.9 and 1.5 SNPs, respectively, and the between-clade mean pairwise distance was 19 SNPs.

Molecular clock analysis and demographic reconstruction

For a more detailed analysis, and to take advantage of the strict serial sampling strategy, we first constructed a maximum parsimony phylogeny from a core set of 237 SNPs, specific to C2. We then used TEMPEST to estimate 'clockiness' of our data[20]. Using linear regression analysis of tip dates versus root-to-tip distances, we confirmed the presence of a strong clock-like signal ($P = 5.5 \times 10^{-12}$) with an average mutation rate of 0.21 per genome per year. The resulting time of most recent common ancestor (tMRCA) was within the first half of 1960 CE (**Figure S1**).

We used BEAST to perform phylogenetic analysis for a more robust estimation of mutation- and growth rate of the outbreak. We evaluated nine different Bayes models and found a constant molecular clock model within an exponentially growing population to be the best fit (**Table S2**). The constant clock model consistently scored higher than relaxed clock models, regardless of the chosen coalescent model. We arrived at a mean mutation rate of 0.24 SNPs per genome per year [95% CI: 0.19 – 0.29], with a tMRCA in late 1959 CE [95% CI: 1944 – 1973]. Thus, the two approaches of estimating the C2 mutation rate were consistent. We also ran the same analyses on a subset of the data in which we adjusted the sampling proportion to 0.1 each year by removing samples from 1992-2000. However, this did not change these and subsequently derived results significantly.

We estimated the historical distance between C2 and SAM5, using an uncorrelated relaxed clock with an underlying lognormal distribution, combined with a bayesian skyline model, similar to Kay et al[15]. Using this approach on SNPs defined by C2, SAM5 and H37Rv (**Figure S2**), we estimated a tMRCA between C2 and the SAM5 set to 1700 CE [95% CI: 1597 - 1792].

From the exponential growth model, we derived a mean posterior growth rate of 0.0744 [95% CI: 0.0398 – 0.111], corresponding to a doubling time of 9.4 years. A non-parametric Skygrid model[21], which ranked second, followed an almost identical trajectory from the tMRCA event until 2005, but displayed a slowing of growth followed by a slow declining trend (**Figure 2a**). The Skygrid model also arrived at a slightly younger tMRCA (1961) than the exponential growth model.

Despite the relatively large credibility interval, the posterior mean growth rate is consistent with corresponding MIRU typing data, in which Danish 1112-15 typed isolates can be fitted to an exponential growth trend at a rate of 0.0706 ($R^2 = 0.736$, $P = 1.6 \times 10^{-7}$) against an overall slight decline in total TB cases in Denmark (exponential growth rate: -0.0148, P = 0.0015) over the same period (**Figure 2b**).

Clustering analysis and identification of robust SNP-markers

From the set of generated BEAST trees, we constructed a timed consensus phylogeny to describe the C2 set from a more epidemiological perspective (**Figure 3a**). This tree was found to be overall congruent with the previously constructed maximum parsimony tree (**Figure S1**), apart from branch lengths, where BEAST incorporates tip dates. The C2-Mj

and C2-Mn clades had tMRCAs of 1966 CE and 2000 CE, respectively, meaning that the former emerged about 34 years before the latter.

To infer epidemiological links between C2 isolates for which we currently possess whole genome data, we derived a clustering threshold of 7 SNPs from the maximum number of possible SNPs over the 23-year sampling period, given the upper credibility limit of the BEAST-estimated mutation rate. From this analysis, we could then define three major epidemic groups of putatively linked isolates, comprising 89% (Group A; 102 cases), 7% (Group B; 8 cases) and 2.6 % (Group C; 3 cases) of C2 cases, respectively. Groups A and C lie within the C2-Mj clade while group B engulfs the whole C2-Mn clade. The clustering was therefore congruent with the underlying phylogeny. Additionally, almost all (100/102) Group A-cases contained a single conserved non-synonymous mutation in the gene *tcrY* (Rv3764c).

From the consensus tree, we were able to distinguish eight smaller strongly supported sub-groups (A.1 – A.7 and B.1), each consisting of five or more isolates. Within each of these sub-groups, we used the inferred median root node age to estimate the most likely year of their emergence. We found that six out of eight of the smaller epidemic groups originated around or after 1992 (**Figure 3b**). Thus, in all likelihood, the IRLM strain collection possess complete transmission chains for these groups, and would benefit from deeper sampling to elucidate patient-to-patient transmission. For the benefit of future studies of the C2 outbreak, we therefore devised a SNP-barcode, comprising fourteen SNP positions from which to classify isolates into each of the currently eight defined epidemic groups (**Figure 4**). Where possible, we selected synonymous variants, to minimize the effect of natural selection, like Coll et al.[16]

Transmission dynamics before and during sampling period

We used TRANSPHYLO[22,23] to infer transmission dynamics during- and prior to the sampling period, based on the maximum clade credibility tree produced by BEAST (**Figure S3**). We used prior distributions for generation- and sampling times (time between infection and transmission, and between infection and sampling, respectively) from a previously analyzed TB outbreak[24] (**Figure S4**). From the consensus transmission tree, we could then derive the likely number of transmission events each year, dating back from 2014 (**Figure 3c**). According to this analysis, the earliest transmission event leading to the current C2 outbreak occurred during 1966, in accordance with the emergence of the C2-Mj clade. The transmission-frequency declined toward the end of the sampling period, consistent with a median lag time between transmission and progression to active disease of one to two years.

The cumulative number of estimated transmission events prior to 1992 was 119, while the total number of transmission events (1966-2014), leading to the 115 sampled cases, amounted to 498. Thus, according to our transmission analysis, roughly one quarter of total C2 transmissions occurred before systematic collection and typing of TB isolates began at the SSI, over a period similar in length (26 years) to the sampling period.

In TRANSPHYLO, we were also able to extract a number of parameters from posterior distributions that describe overall characteristics of the C2 outbreak (**Figure S5**). The mean basic reproductive number R_0, defined as the expected number of secondary diseases per case in a completely susceptible population, was 1.15 [95% CI: 1.08 – 1.23], which is consistent with a moderately expanding TB outbreak ($R_0 > 1$) in a low-burden setting[22]. The mean within-host coalescence rate ($N_e g$) was 1.33 [95% CI: 0.53 – 3.46] events per year, corresponding to a within-host effective population size of 485 [95% CI: 193 – 1263]. The mean sampling proportion (π) was 0.05 [95% CI: 0.04 - 0.07], which was somewhat lower than the mean sampling proportion of typed isolates over the sampling period (0.116, from 115/989 isolates; see also **Figure 2b**).

Discussion

To characterize the largest ongoing clonal TB outbreak (C2) in Scandinavia, we subjected a representative time-series of 115 *Mtb* isolates collected 1992-2014, to WGS. The restrained sampling strategy reflects the immense size of this cluster, which by now has exceeded one thousand cases.

Our initial observation from phylogenetic analysis is that C2 consists of two discernible clades (Figure 3), one major (C2-Mj) and one minor (C2-Mn), where C2-Mj is the older and more diverse of the two. Using molecular dating we arrived at 1959 CE [95% CI: 1944 – 1973] as the year of initial diversification, pointing to introduction of C2 into Denmark sometime after the Second World War. Thus, we estimate that emergence, or possibly, the re-emergence, of C2 two to five decades prior to the beginning of routine collection and genotyping of Danish TB isolates in 1992.

We calculated an overall mutation rate of 0.24 SNPs [95% CI: 0.19 – 0.29] per genome per year for C2, which agrees well with other studies, typically reporting between 0.2 - 0.5 SNPs per genome per year[25–28].

We defined three distinct epidemic groups (A, B and C) using a clustering threshold of 7 SNPs, which we derived from the estimated mutation rate over the 23-year sampling period. Thus, even if the underlying sample set was initially defined by a single MIRU-VNTR genotype (1112-15), our WGS analysis reveals several distinct but historically linked transmission chains, even without directly inferring patient-to-patient transmissions. Incidentally, if the clustering threshold was increased to include the initial year of diversification (1959), the resulting threshold, 16 SNPs over 55 years, produced a single cluster containing all 115 C2 isolates.

The largest of the epidemic groups (Group A) contains 89% (102/115) of the sequenced isolates with a consensus tMRCA in 1975, which predates routine typing at the IRLM by 14 years. We therefore further subdivided Group A into groups consisting of five or more isolates with strong clade support (A.1 – A.7). Using the same criteria, we could also identify a single subgroup (B.1) inside epidemic group B.

Even though the C2 dataset consists mainly (108/115) of samples collected over the entire greater Copenhagen area, we observed that two of the epidemic groups (subgroup A.3 and group B) could be specifically linked to discrete geographic areas within the capital region. In subgroup A.3, for example, 10 out of 11 strains stemmed from the same postal district with a population size of 36.000, while group B contained 6 out of 8 cases stemming from an area around a shelter/drop-in center in Copenhagen. This agrees with previous studies, in which subclades have been linked epidemiologically[5,8,19], strengthening the utility of WGS-based subtyping for transmission tracing, especially in lieu of detailed epidemiological contact data. Each of the eight subgroups can be identified through the presence of one or more conserved SNPs, from which we have selected a core set of fourteen for future classification of C2 isolates (See Table 1).

Previously it has been shown that different MIRU-types can be linked into contiguous transmission chains with WGS[19]. Consequently, it is likely that more cases could be included in the C2 outbreak if related MIRU-VNTR types from the IRLM collection were investigated. Our collection currently holds 45 isolates from 19 different genotypes that differ from 1112-15 in only one locus. Isolates from three of these genotypes have so far been confirmed to extend from the C2 outbreak using the presented core SNP set (unpublished).

The deep branching of the C2-Mn clade (Group B), which we estimated to originate in 2000, points to a possible case of reactivation following prolonged latency. Despite the lack of a clear index case and the relatively sparse sampling of C2, we deem it unlikely that deeper or complete sampling would produce numerous intermediates to challenge this. Thus, we find it likely that the nine common SNPs observed in this deep branch accumulated during latent infection. In light of prior observations, that mutation accumulation can occur during prolonged latency at rates comparable to active disease[13], we underpin that latent cases continually pose a threat to the emergence of new transmission chains[29].

Our analysis also established that the C2 outbreak belongs to MTBC sublineage L4.8. Recently, a major study of the Euro-American MTBC Lineage 4 described the newly defined L4.10/PGG3 sublineage (combining L4.7, L4.8 and L4.9), as a 'generalist' with widespread global distribution[30]. Given that C2 is the largest epidemic cluster in Scandinavia, it is unsurprising to find it linked to such a ubiquitous sublineage. The most closely related isolates specifically originated from Russia, and we estimated that C2 is separated from its closest Russian isolate by at least 200 years. This would suggest Russia or possibly Eastern Europe as the historical origin of C2. Interestingly, the estimated emergence of C2 in Denmark coincides with the influx of approximately 1,000 Hungarian immigrants in the wake of the Hungarian Revolution of 1956. A recent study by Eldholm et al, exemplified the impact of political instability on the spread and diversification of TB[31]. Unfortunately, apart from a minor collection of *Mtb* strains from the 1960s, which was not found to contain the C2-genotype[32], the Danish TB collection does not hold isolates from before 1992, making it difficult to delve deeper into this hypothesis at this point.

In conclusion, using WGS on only 12% of culture-verified isolates collected over 23 years, we have derived key information about the largest ongoing TB-outbreak in Scandinavia that was previously only identified as a single genotype. It remains unclear whether the success of C2 is attributable to a unique virulence profile but we have so far not found evidence to suggest this. Moreover, the extent of the C2 outbreak in Denmark has previously been linked to delayed diagnosis[1,2]. Regardless, our present study provides a strong case for strengthening TB control in high-risk groups to curtail active transmission and monitor re-activation of latent cases. Consequently, we continue to stress the need for sustained TB awareness, even in low-incidence settings such as Denmark. In Copenhagen, efforts to screen for TB by sputum culture in high-risk groups has been ongoing since 2012[33]. Identification is also important for proper control actions, and with this study we provide a simple example of using a SNP-typing scheme,

derived from WGS analysis of historical isolates, for rapid classification of discrete transmission chains in large ongoing outbreaks.

References

1. Bauer, J., Yang, Z., Poulsen, S. & Andersen, A. B. Results from 5 years of nationwide DNA fingerprinting of *Mycobacterium tuberculosis* complex isolates in a country with a low incidence of *M. tuberculosis* infection. J Clin Microbiol **1998** ; 36**:**305–8.

2. Lillebaek, T., Dirksen, a, Kok-Jensen, a & Andersen, a B. A dominant *Mycobacterium tuberculosis* strain emerging in Denmark. Int J Tuberc Lung Dis **2004** ; 8:1001–6.

3. Kamper-Jørgensen, Z. *et al.* Clustered tuberculosis in a low-burden country: nationwide genotyping through 15 years. J Clin Microbiol **2012** ; 50:2660–7.

4. Lillebaek, T. *et al. Mycobacterium tuberculosis* outbreak strain of Danish origin spreading at worrying rates among greenland-born persons in Denmark and Greenland. J Clin Microbiol **2013** ; 51:4040–4.

5. Roetzer, A. *et al.* Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **2013** ; 10:e1001387.

6. Gardy, J. L. *et al.* Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N Engl J Med **2011** ; 364:730–739.

7. Walker, T. M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **2013** ; 13:137–46.

8. Mehaffy, C. *et al.* Marked Microevolution of a Unique *Mycobacterium tuberculosis* Strain in 17 Years of Ongoing Transmission in a High Risk Population. PLoS One **2014** ; 9:e112928.

9. EuroTB. Surveillance of Tuberculosis in Europe. Rep. Tuberc. cases Notif. 2005 **2007**.

10. European Centre for Disease Prevention and Control/WHO. Tuberculosis surveillance in Europe 2009. Reproduction **2011** ; doi:10.2900/28358

11. European Centre for Disease Prevention and Control & WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2015 **2015** ; doi:10.2900/666960

12. Svensson, E. *et al. Mycobacterium chimaera* in Heater – Cooler the United States and United Kingdom. Emerg Infect Dis **2017** ; 23:507–509.

13. Lillebaek, T. *et al.* Substantial molecular evolution and mutation rates in prolonged latent *Mycobacterium tuberculosis* infection in humans. J Med Microbiol **2016** ; doi:10.1016/j.ijmm.2016.05.017

14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **2014** ; 30:2114–2120.

15. Kay, G. L. *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat Commun **2015** ; 6:6717.

16. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun **2014** ; 5:4812.

17. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet **2013** ; 45:1176–1182.

18. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet **2014** ;46, 279–86.

19. Bjorn-Mortensen, K. *et al.* Tracing *Mycobacterium tuberculosis* transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Sci Rep **2016** ; 6:33180.

20. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol **2016** ; 2:vew007.

21. Gill, M. S. *et al.* Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. Mol Biol Evol **2013** ; 30:713–724.

22. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol Biol Evol **2014** ; 31:1869–1879.

23. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. bioRxiv **2016** ; 34:65334.

24. Hatherell, H.-A. *et al.* Declaring a tuberculosis outbreak over with genomic epidemiology. Microb Genomics **2016** ; 1:10.1099/mgen.0.000060.

25. Guerra-Assuncao, J. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife **2015** ; 2015:1–17.

26. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. Nat Commun **2015** ;6:1–9.

27. Bryant, J. *et al.* Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis **2013** ; 13:110.

28. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nat Genet **2011** ; 43:482–6.

29. Sandgren, A. *et al.* Identifying components for programmatic latent tuberculosis infection control in the European Union. Euro Surveill **2016** ; 21:1–5.

30. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet **2016** ; 48:1535–1543.

31. Eldholm, V. *et al.* Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. Proc Natl Acad Sci **2016** ; doi:10.1073/pnas.1611283113

32. Lillebaek, T. *et al.* Stability of DNA patterns and evidence of *Mycobacterium tuberculosis* reactivation occurring decades after the initial infection. J Infect Dis **2003** ; 188:1032–1039.

33. Jensen, S. G. *et al.* Screening for TB by sputum culture in high-risk groups in Copenhagen, Denmark: a novel and promising approach. Thorax **2015** ; 70:979–83.

Conflicts of interest

Dorte Bek Folkvardsen: None

Anders Norman: None

Åse Bengård Andersen: None

Erik Michael Rasmussen: None

Lars Jelsbak: None

Troels Lillebaek: None

Funding

This work was supported by the Novo Nordisk Foundation [grant number NNF7651] and the Lundbeck Foundation [grant number R151-2013-14628]. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

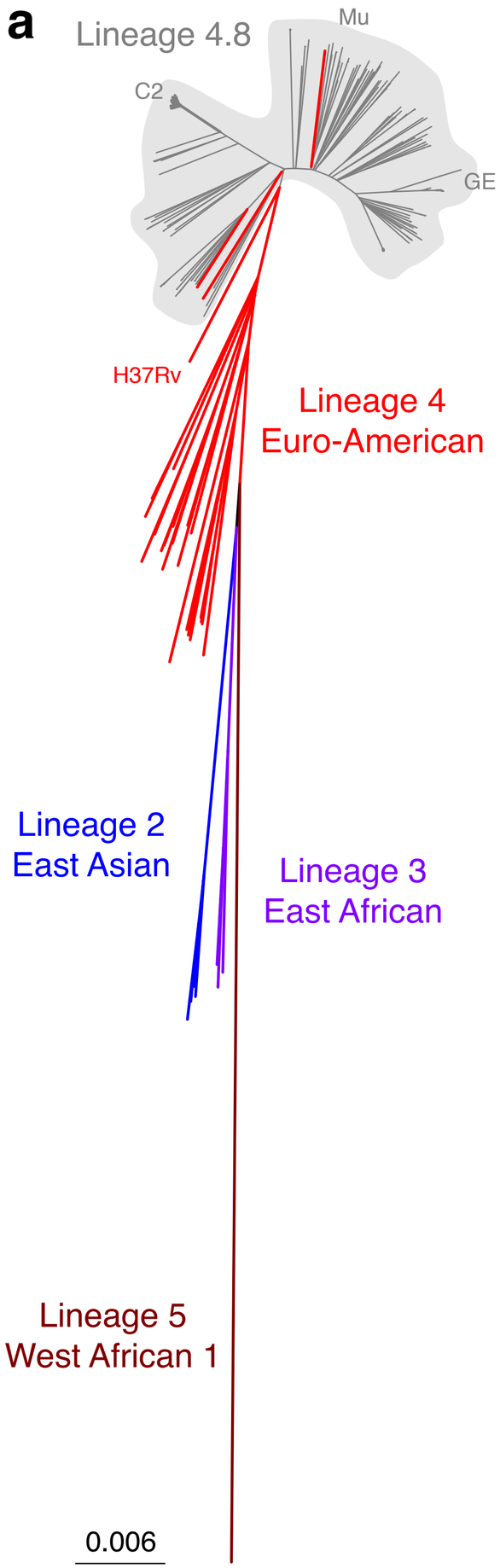
Figure Legends

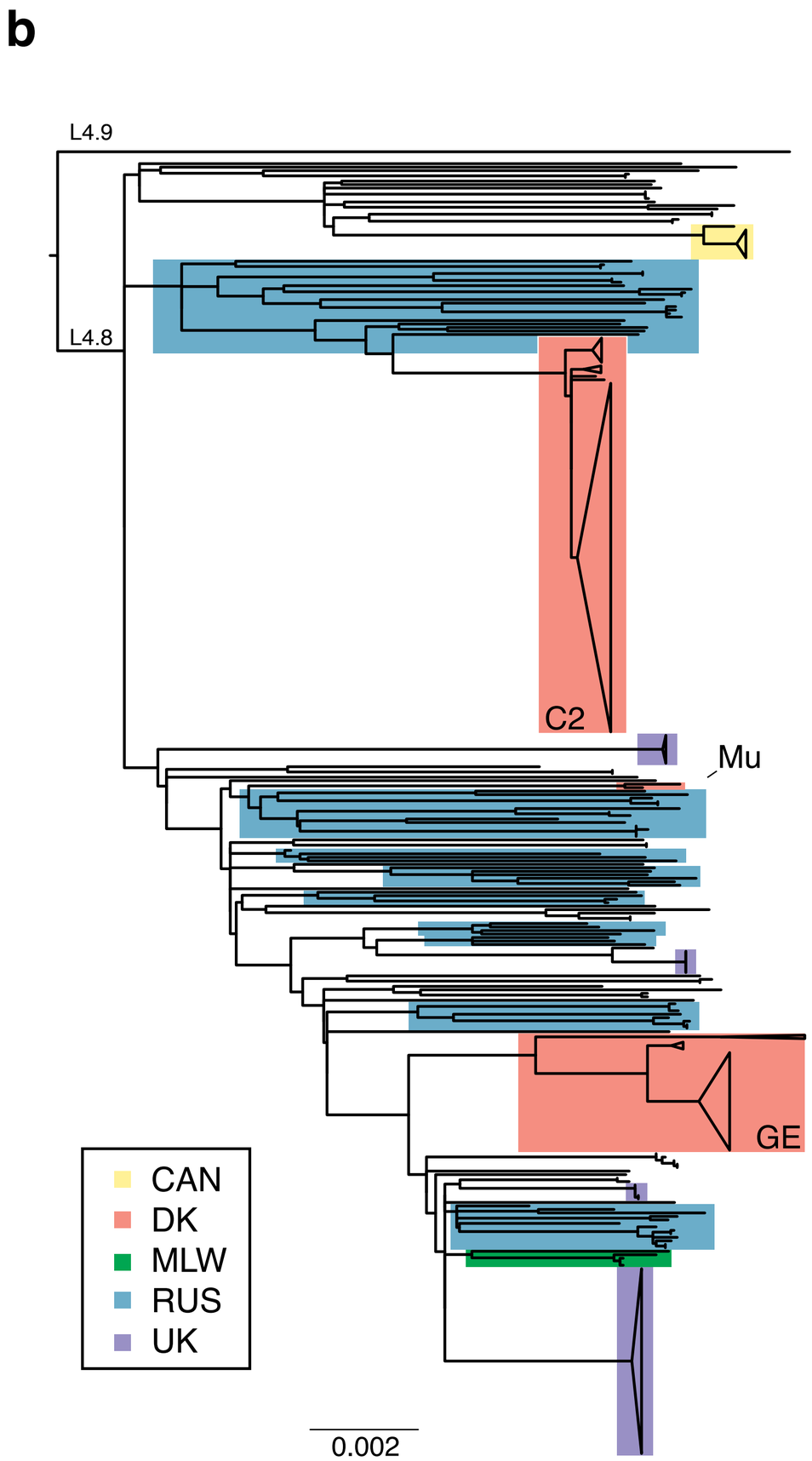
Figure 1. Phylogenetic placement of the Danish Cluster-2 (C2) *M. tuberculosis* outbreak based on single-nucleotide polymorphisms (SNPs) of 115 representative C2-isolates in (a) relation to 33 global MTBC reference strains representing the three "modern" MTBC lineages 2-4 (Comas et al. 2013) and (b) in relation to 254 other TB isolates belonging to the L4.8 sublineage (indicated as gray in panel a). Large clades where all isolates originate from a single country are colored according to the legend (CAN: Canada; DK: Danish Kingdom; MLW: Malawi; RUS: Russia; UK: United Kingdom). The pairwise number of SNPs between C2 isolates and the rest of L4.8 (c), and between just the C2 isolates (d) are shown as histograms. See main text for explanation of the isolate sets GE, Mu and SAM5.

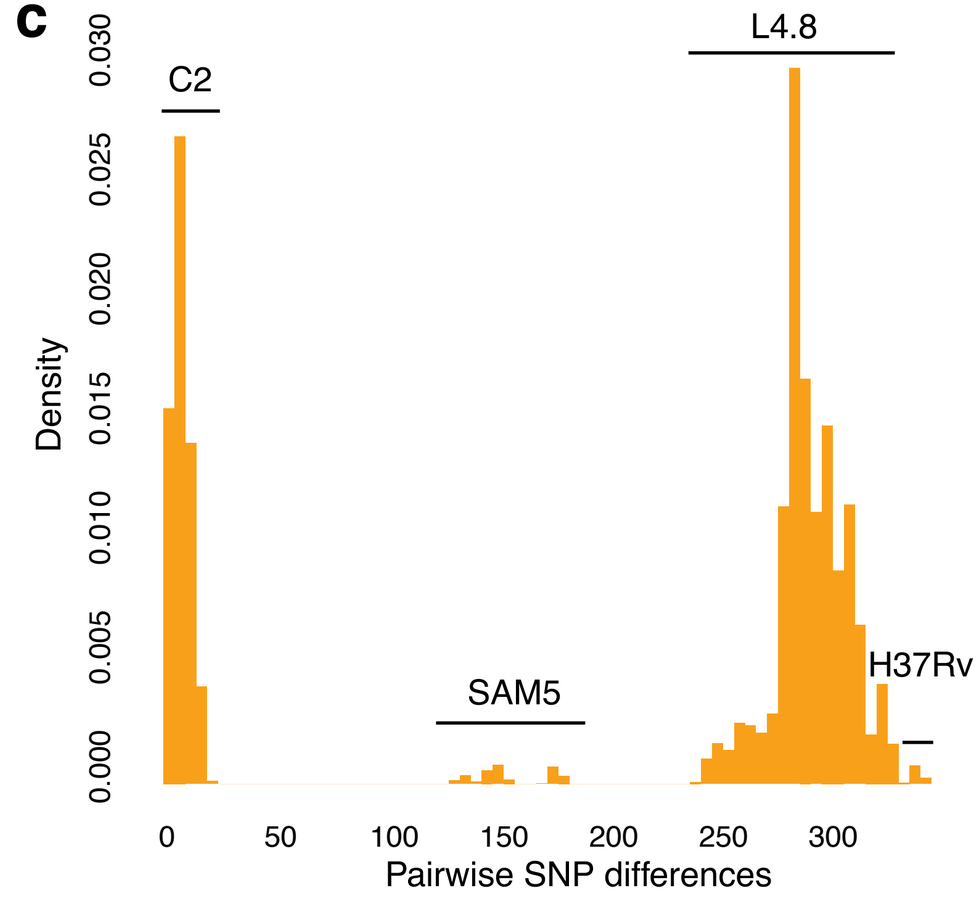
Figure 2. (a) Effective population size of the C2 outbreak over time, including 95% confidence intervals, derived from BEAST-analysis based on exponential growth (black line) and Bayesian Skygrid (blue lines) coalescent models. Vertical dotted lines on the graph represent the estimated time of most recent common ancestor (tMRCA) according to the chosen Bayesian model. (b) The log-number of MIRU-typed isolates of either C2 isolates (red dots) or the total number of isolates from the Danish Kingdom (blue circles) fitted to linear regression lines, showing the overall growth trends. The black circles indicate the sampling proportion out of total C2 isolates, when picking five isolates each year.

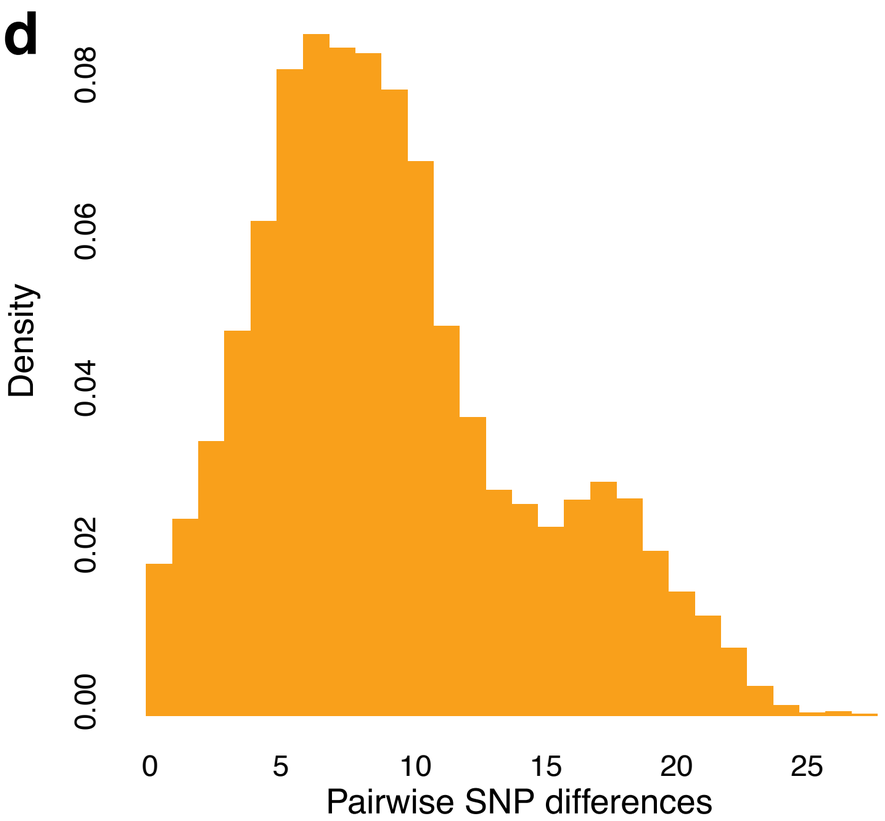
Figure 3. Timed phylogeny based on BEAST consensus tree showing the subdivision of C2 into one major (C2-Mj) and one minor (C2-Mn) clade. Highlighted clades A, B and C represents clusters of epidemiologically linked isolates, defined by a threshold of ≤ 7 SNPs. All but two isolates in cluster A also harbor a non-synonymous SNP in the H37Rv gene *tcrY* (Rv3674c). Smaller epi-groups A.1-A7 and B.1 are defined by clades of five or more isolates with strong clade support (> 70%) on the consensus tree. Horizontal bars (b) indicate the timeline of each epidemic group from the estimated time of most recent common ancestor (light colored areas) and the period in which representative isolates were collected (dark colored areas). The number of transmission events per year, inferred from TRANSPHYLO analysis, is shown along the timeline as a histogram (c). The vertical black dotted line indicates the year (1992) in which systematic typing of TB isolates began at the IRLM.

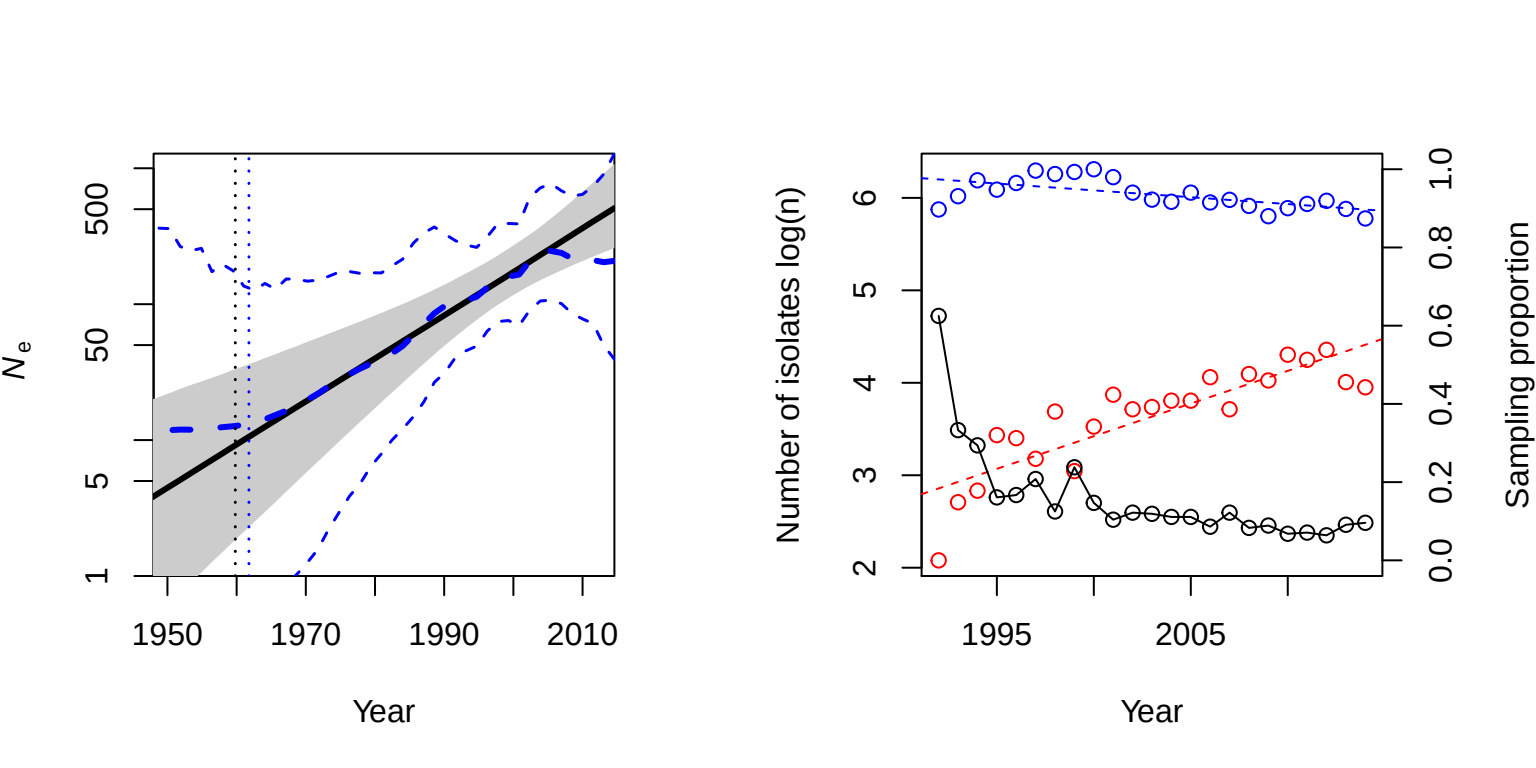
Table 1. Overview of core SNPs-set useful for identifying C2 and its major (C2-Mj) and minor (C2-Mn) clades, the three epidemic groups A,B & C of epidemiologically linked isolates, and eight smaller epidemic groups that are part of the C2 outbreak.

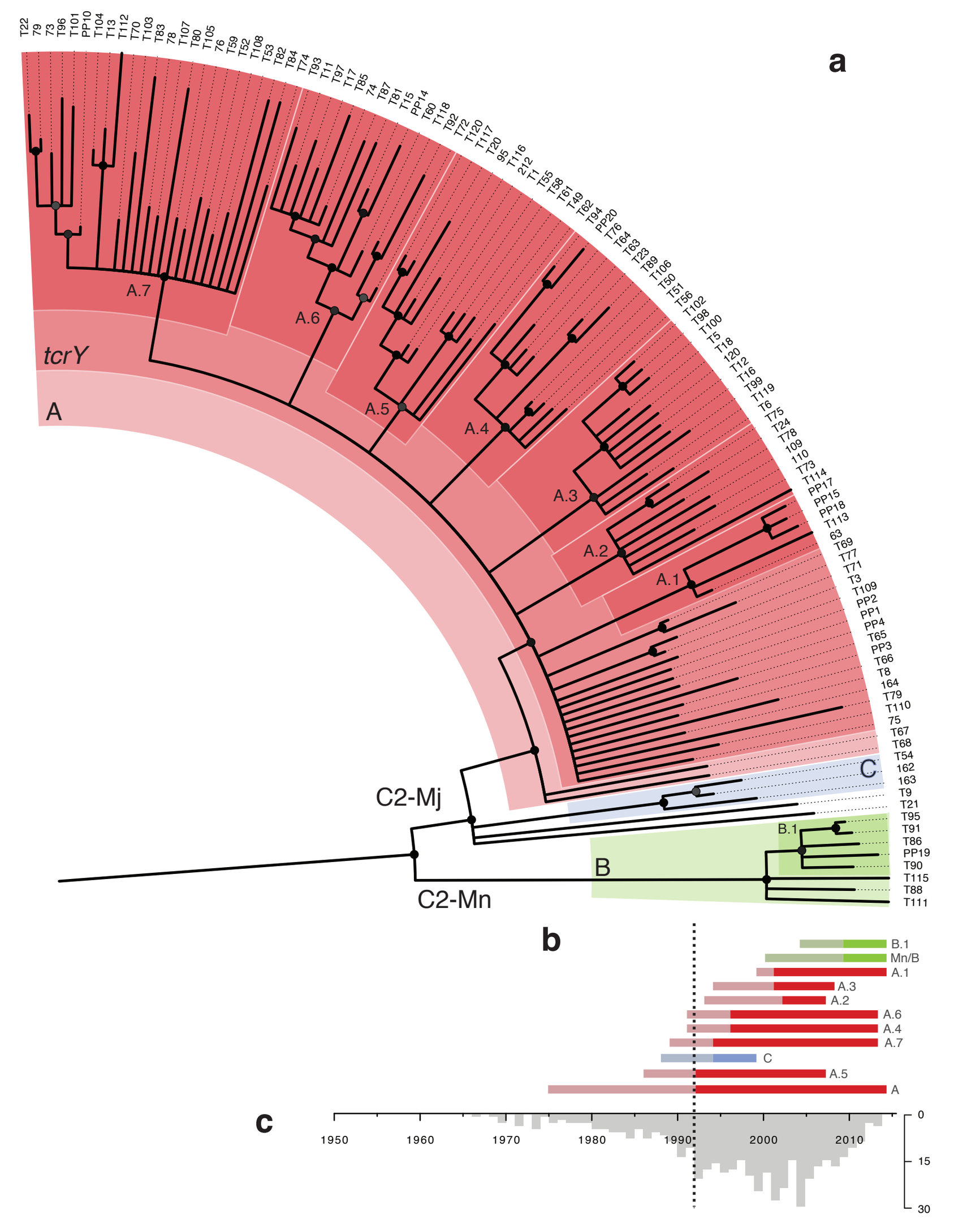

a
Lineage 4.8
Mu
C2
GE
H37Rv
Lineage 4
Euro-American
Lineage 2
East Asian
Lineage 3
East African
Lineage 5
West African 1
0.006


b
L4.9
L4.8
C2
Mu
GE
CAN
DK
MLW
RUS
UK
0.002


c
C2
L4.8
SAM5
H37Rv
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
50
100
150
200
250
300
Pairwise SNP differences


d
Density
0.00
0.02
0.04
0.06
0.08
0
5
10
15
20
25
Pairwise SNP differences


N_e
1
5
50
500
1950
1970
1990
2010
Year
Number of isolates log(n)
2
3
4
5
6
1995
2005
Year
0.0
0.2
0.4
0.6
0.8
1.0
Sampling proportion


a
T22
79
73
T96
T101
PP10
T104
T13
T112
T70
T103
T83
78
T107
T80
T105
76
T59
T52
T108
T53
T82
T84
T74
T93
T11
T97
T17
T85
74
T87
T81
T15
PP14
T60
T118
T92
T72
T120
T117
T20
95
T116
212
T1
T55
T58
T61
T49
T62
T94
PP20
T76
T64
T63
T23
T89
T106
T50
T51
T56
T102
T98
T100
T5
T18
120
T12
T16
T99
T119
T6
T75
T24
T78
109
110
T73
T114
PP17
PP15
PP18
T113
63
T69
T77
T71
T3
T109
PP2
PP1
PP4
T65
PP3
T66
T8
164
T79
T110
75
T67
T68
T54
162
163
T9
T21
T95
T91
T86
PP19
T90
T115
T88
T111
A.7
A.6
A.5
A.4
A.3
A.2
A.1
tcrY
A
C
B.1
B
C2-Mj
C2-Mn
b
B.1
Mn/B
A.1
A.3
A.2
A.6
A.4
A.7
C
A.5
A
c
1950
1960
1970
1980
1990
2000
2010
0
15
30

Position	Reference	Variant	Locus	NT change	AA change	Annotation	Epi-group	No. conserved SNPs	C2	C2-Mn/B	B.1	C2-Mj	A	TcrY	A.1	A.2	A.3	A.4	A.5	A.6	A.7	C
66,196	G	T	Rv0062	G645T	.	Possible cellulase CelA1	C2	59	T	T	T	T	T	T	T	T	T	T	T	T	T	T
98,168	C	T	Rv0089	C411T	.	Possible methyltransferase/methylase	A.1	5	.	.	.	.	.	.	T	.	.	.	.	.	.	.
164,112	C	T	cyp138 (Rv0136)	C747T	.	Probable cytochrome P450	A.7	3	.	.	.	.	.	.	.	.	.	.	.	.	T	.
240,943	T	C	mmpL11 (Rv0202c)	A351G	Y117C	Probable conserved transmembrane transport protein	A.3	1	.	.	.	.	.	.	.	.	C	.	.	.	.	.
257,844	T	G	fadE3 (Rv0215c)	A1014C	N338T	Probable acyl-CoA dehydrogenase	C2-Mj	2	.	.	.	G	G	G	G	G	G	G	G	G	G	G
324,693	G	A	fadD2 (Rv0270)	G127A	G43R	Probable fatty-acid-CoA ligase	A.6	1	.	.	.	.	.	.	.	.	.	.	.	A	.	.
346,380	C	T	eccC3 (Rv0284)	C746T	A249V	ESX-3 type VII secretion system protein	B.1	1	.	.	T	.	.	.	.	.	.	.	.	.	.	.
472,569	A	G	ndhA (Rv0392c)	T72C	F24S	Probable membrane NADH dehydrogenase			.	.	.	.	.	.	.	.	.	.	.	.	.	G
497,759	T	G	glnH (Rv0411c)	A543C	D181A	Probable glutamine-binding lipoprotein GlnH (GLNBP)	A.2	1	.	.	.	.	.	.	.	G	.	.	.	.	.	.
777,590	C	T	end (Rv0670)	C522T	.	Probable endonuclease IV End	C	4	.	.	.	.	.	.	.	.	.	.	.	.	.	T
997,982	C	G	Rv0895	C201G	Y67*	Possible triacylglycerol synthase			.	.	.	.	.	.	.	.	.	.	.	.	G	.
1,035,840	C	T	prrA (Rv0903c)	T280C	.	Two component response transcriptional regulatory protein			.	.	.	.	T	T	T	T	T	T	T	T	T	.
1,410,569	C	G	amiB2 (Rv1263)	C139G	R47G	Probable amidase AmiB2 (aminohydrolase)			.	.	.	.	.	.	G	.	.	.	.	.	.	.
1,467,344	C	T	atpC (Rv1311)	C30T	.	Probable ATP synthase epsilon chain	A	2	.	.	.	.	T	T	T	T	T	T	T	T	T	.
1,813,265	C	T	lgt (Rv1614)	C95T	T32I	Possible prolipoprotein diacylglyceryl transferases			.	T	T	.	.	.	.	.	.	.	.	.	.	.
1,835,148	C	T	coaE (Rv1631)	C136T	.	Probable dephospho-CoA kinase	C2-Mn/B	9	.	T	T	.	.	.	.	.	.	.	.	.	.	.
1,951,980	C	T	Rv1726	C129T	.	Probable oxidoreductase			.	T	T	.	.	.	.	.	.	.	.	.	.	.
2,242,582	G	A	ctpF (Rv1997)	G2424A	.	Probable metal cation transporter P-type ATPase A	A.5	1	.	.	.	.	.	.	.	.	.	.	A	.	.	.
2,329,920	G	A	cobL (Rv2072c)	C228T	P76L	Precorrin-6Y C(5,15)-methyltransferase	A.4	1	.	.	.	.	.	.	.	.	.	A	.	.	.	.
2,528,014	G	A	Rv2253	G31A	A11T	Possible secreted unknown protein			.	.	.	.	.	.	.	.	.	.	.	.	.	A
2,537,023	A	G	Rv2264c	T1329C	V443A	Conserved hypothetical proline rich protein			.	G	G	.	.	.	.	.	.	.	.	.	.	.
2,563,075	G	A	lppO (Rv2290)	G477A	M159I	Probable conserved lipoprotein			.	A	A	.	.	.	.	.	.	.	.	.	.	.
2,603,748	G	A	lppP (Rv2330c)	C476T	H159Y	Probable lipoprotein			.	A	A	.	.	.	.	.	.	.	.	.	.	.
2,768,599	T	C	Rv2466c	A287G	N96D	Conserved protein			.	.	.	.	.	.	C	.	.	.	.	.	.	.
3,290,947	T	G	lppX (Rv2945c)	A380C	S127R	Probable conserved lipoprotein			.	.	.	.	.	.	.	.	.	.	.	.	G	.
3,405,449	C	T	fecB (Rv3044)	C314T	A105V	Probable FEIII-dicitrate-binding periplasmic lipoprotein			.	.	.	.	.	.	.	.	.	.	.	.	.	T
3,600,762	C	T	Rv3224A	C128T	P43L	Conserved hypothetical protein			.	T	T	.	.	.	.	.	.	.	.	.	.	.
3,805,888	G	A	acrA1 (Rv3391)	G268A	A90T	Possible multi-functional enzyme with acyl-CoA-reductase activity			.	A	A	.	.	.	.	.	.	.	.	.	.	.
3,904,014	C	T	cpsA (Rv3484)	C937T	.	Possible conserved protein			.	T	T	.	.	.	.	.	.	.	.	.	.	.
4,209,859	C	A	tcrY (Rv3764c)	G1152T	R384L	Possible two component sensor kinase	TcrY	1	.	.	.	.	.	A	A	A	A	A	A	A	A	.